REGRESSION PART 6: EXAMPLE

Hsin-Min Lu
盧信銘
台大資管系

Case Study:
Predicting Releasing Year of a Song

- Dataset: Million song dataset from UCI
- http://archive.ics.uci.edu/ml/datasets/YearPredictionMSD

- Training: 463,715 songs
- Testing: 51,630 songs

- The default train-test split “avoids the 'producer effect' by making sure no song from a given artist ends up in both the train and test set.”
- 90 attributes, 12 = timbre average, 78 = timbre covariance.
- 'timbre' features extracted using the Echo Nest API.

Prediction Models and Goals

- Prediction Models:
  - Ordinary Linear Regression
  - Ridge Regression
  - K-nearest Neighbors (kNN) Regressions

- Goals:
  - Understand the effect of sample size on prediction power for
    (a) a given model, and
    (b) across different models
  - Understand the effect of feature weighting on kNN regression.

More about Prediction Models

- OLS (referred to as Linear Regression): Use the standard close form solution for regression coefficients.
- Ridge Regression (Linear Models with L2 regularization).
- $\min_w \|\boldsymbol{X}w - \boldsymbol{t}\|_2^2 + \alpha\|w\|_2^2$
- Use cross validation to select best α.
- Grids: 0.01 to N/3, geometric progression. 20 grids.
-
- kNN: Use Euclidian distance ($\|x_1 - x_2\|_2$) to determine neighbors.
- Will not select k using training data
- Will show prediction performance on different k.

Feature Preprocessing

- kNN is sensitive to feature scaling because of the Euclidian distance metric adopted.
- Solution: Standardize features to have mean = 0 and variance = 1.
- Meaning: all features are equally important.

- Note: to be fair, all models will use the standardized feature value.

- Can we do better?

Feature Weighting and Prediction Performance Measure

- Use regression coefficients from ridge regression to weight feature values.

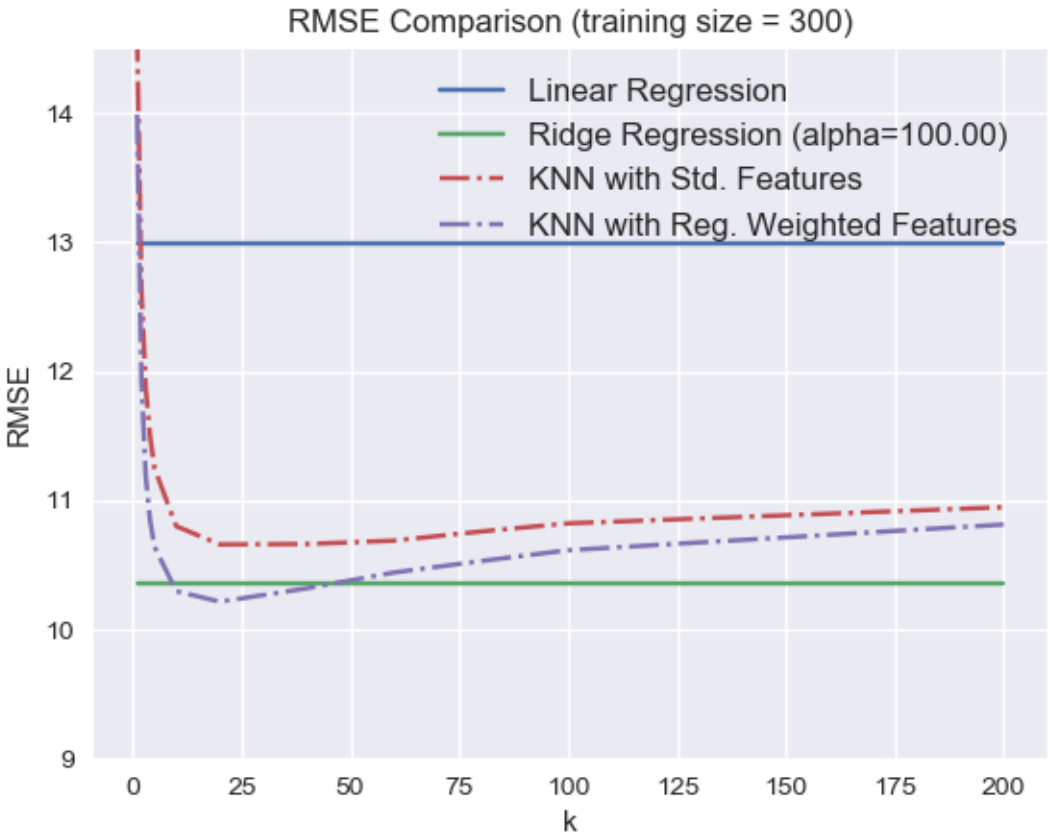
- Reason: Higher regression coefficients suggest higher importance to predict the outcome.

- Prediction Performance Measure: RMSE

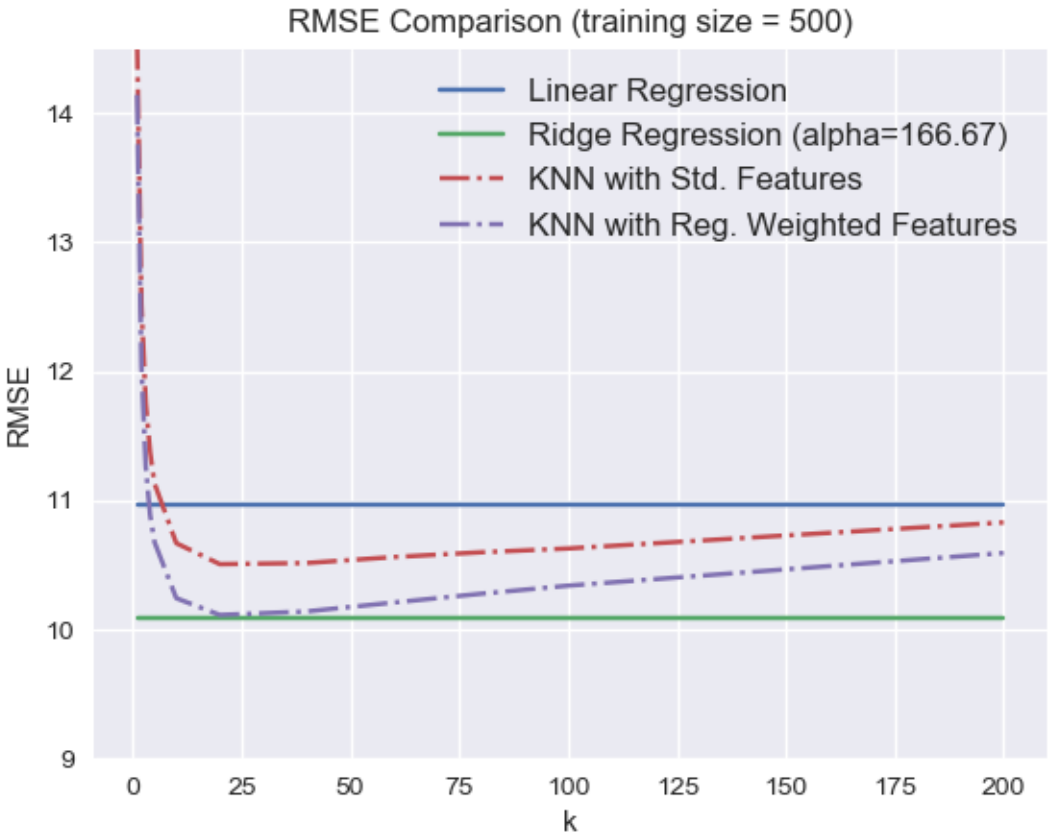
- RMSE = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - \hat{t}_i)^2}$

Geographic Distribution of Songs

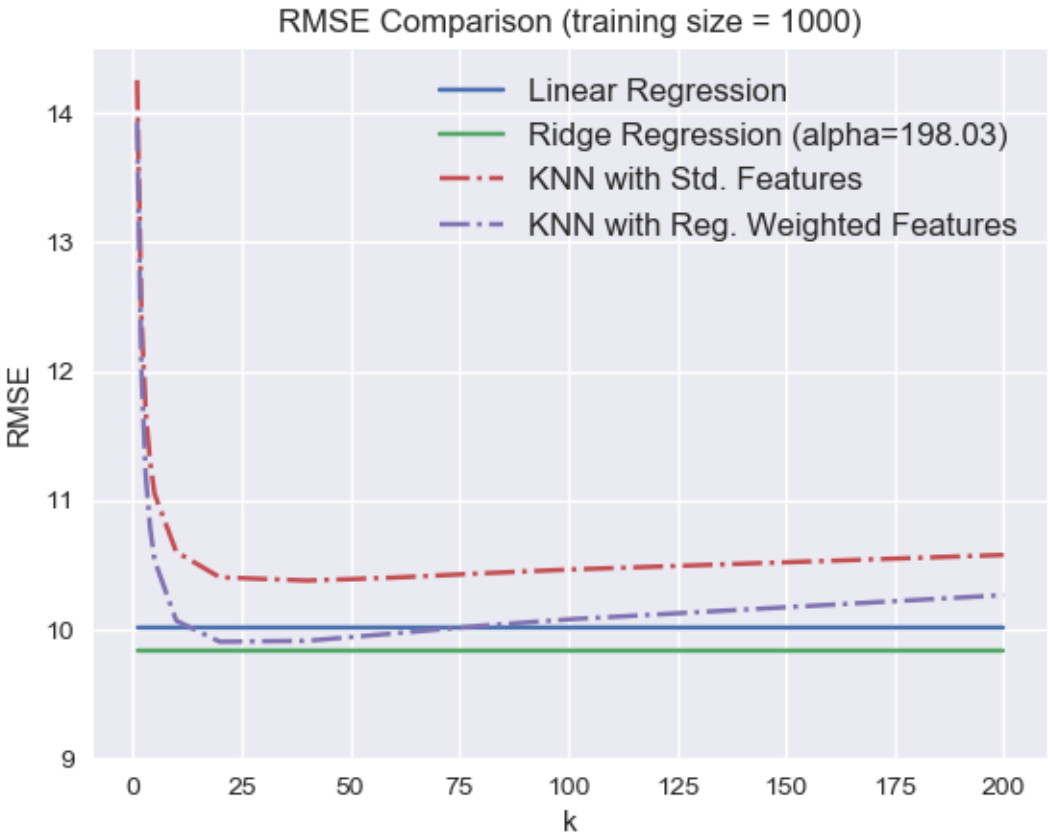
When Data is Scares (training = 300)



Training = 500



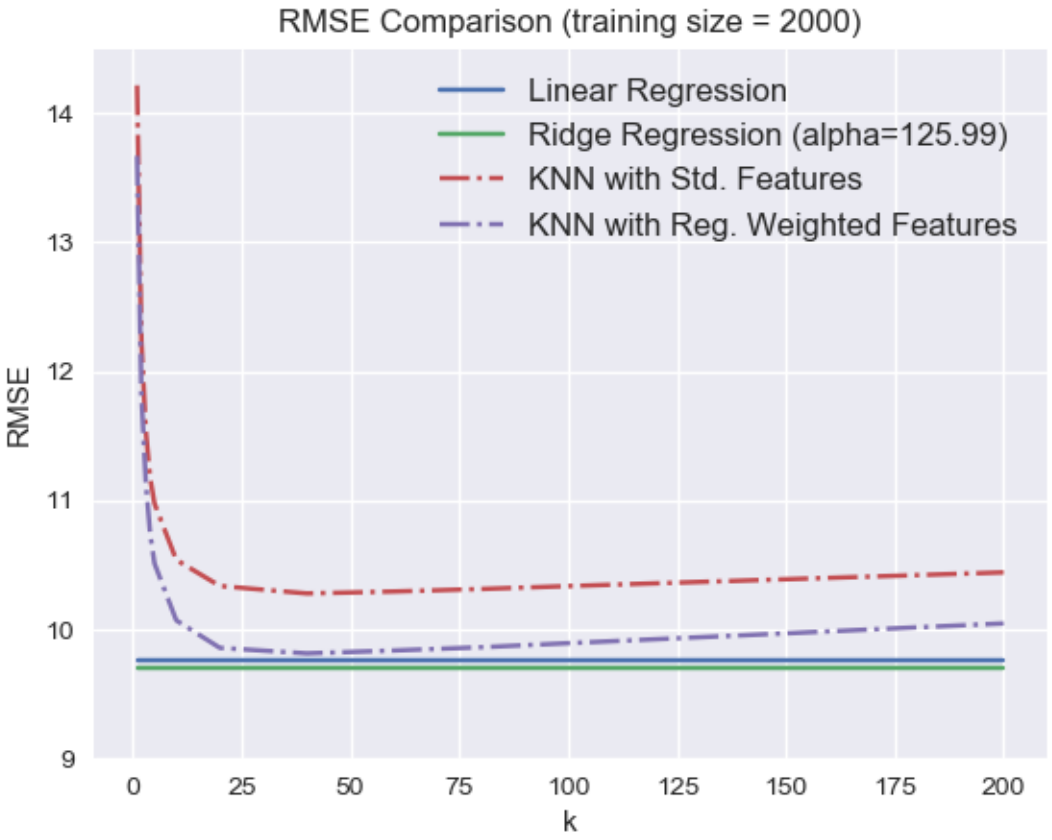
Training = 1000



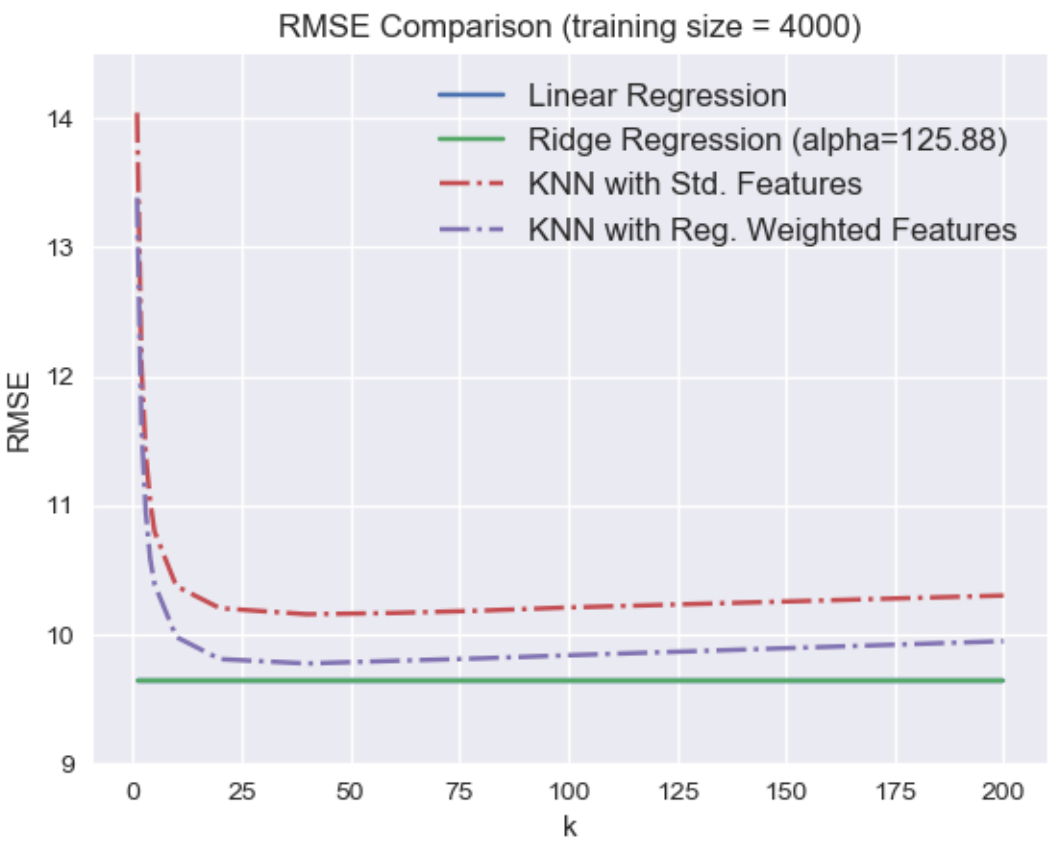
Observation

- Regularization is very effective when the training data size is small.
- Larger sample size benefits all approaches.
- Linear regression improved dramatically from n=300 to n=1000.
- Ridge regression also benefits from larger training set. But the gap is much smaller compared to linear regression.
- kNN with weighted features always performed better than using standardized features.

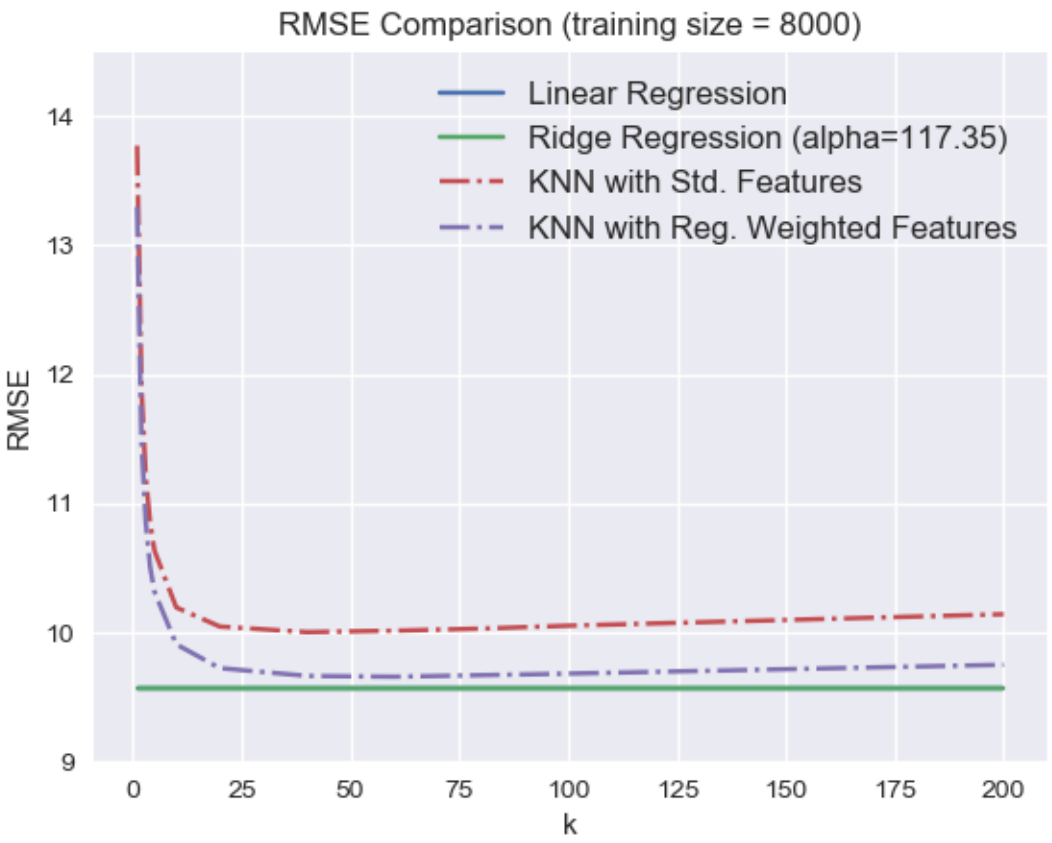
Training = 2000



Training = 4000



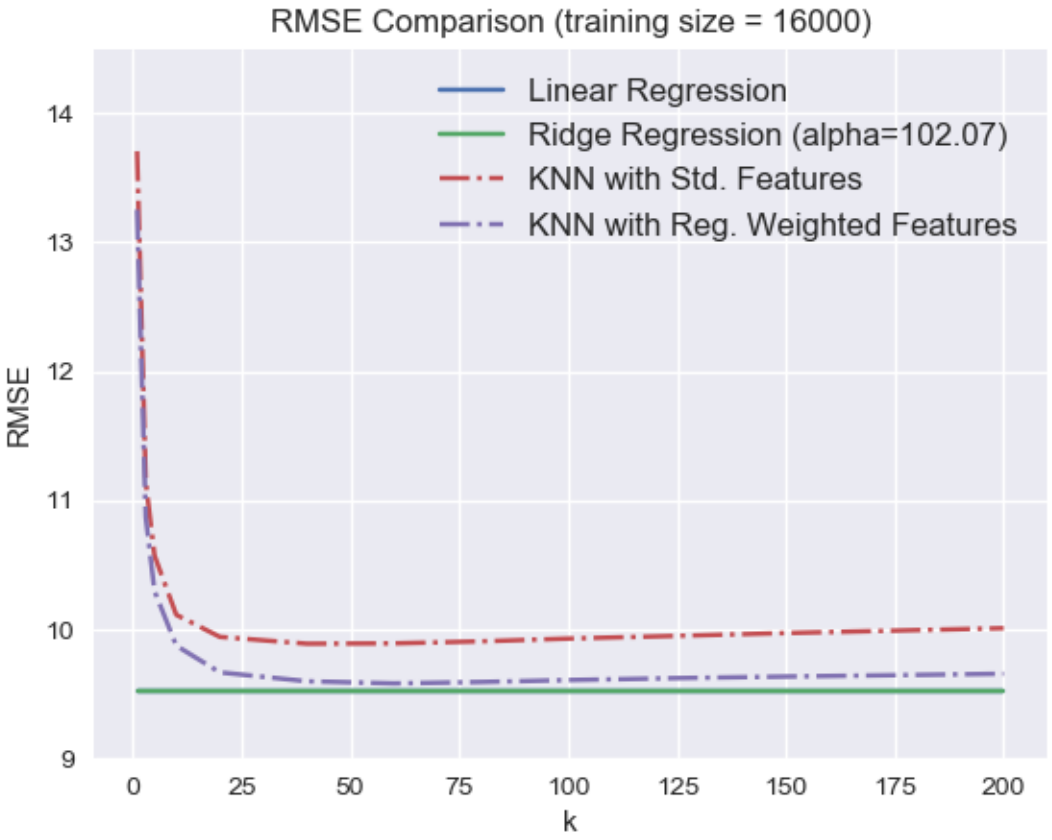
Training = 8000



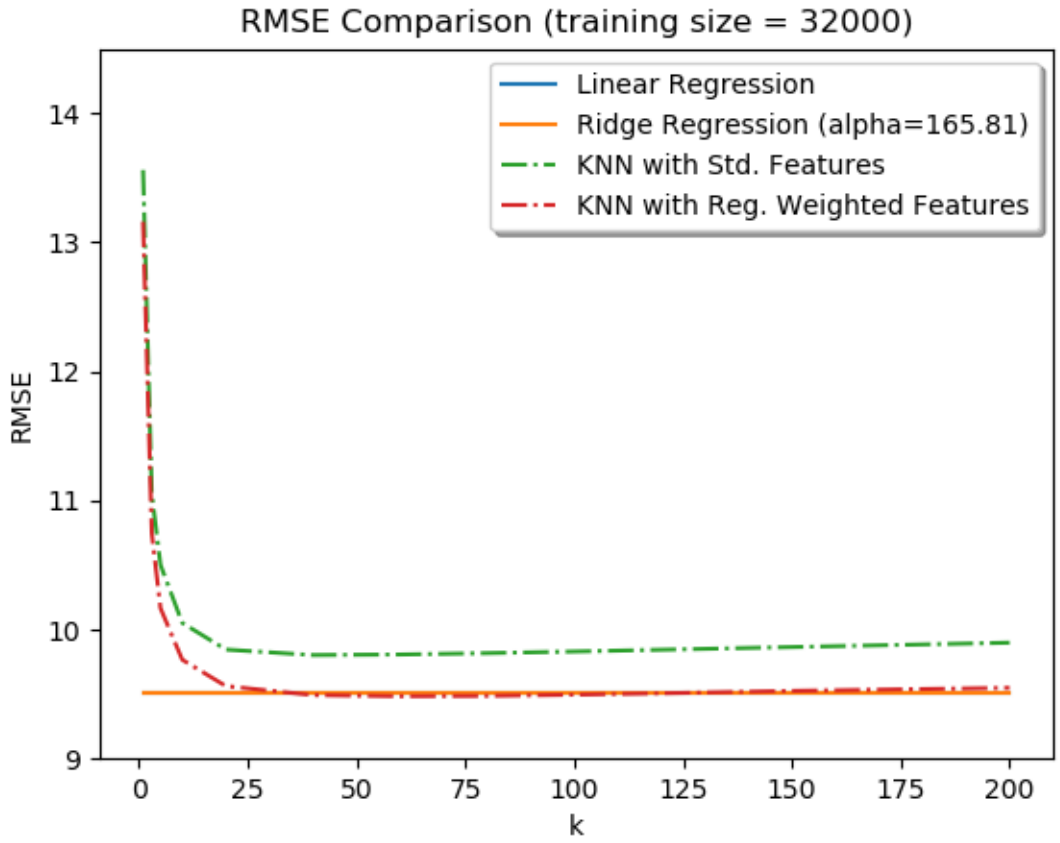
Observations (Part 2)

- For larger training dataset, the effect of regularization is small.
  - Why? Is there situations when regularization is important for large training dataset?
- Increasing training dataset still benefit all prediction models.
  - The improvements, however, slows down.
- Interestingly, kNN is not better than linear regression or ridge regression.
  - Why? Any guess?

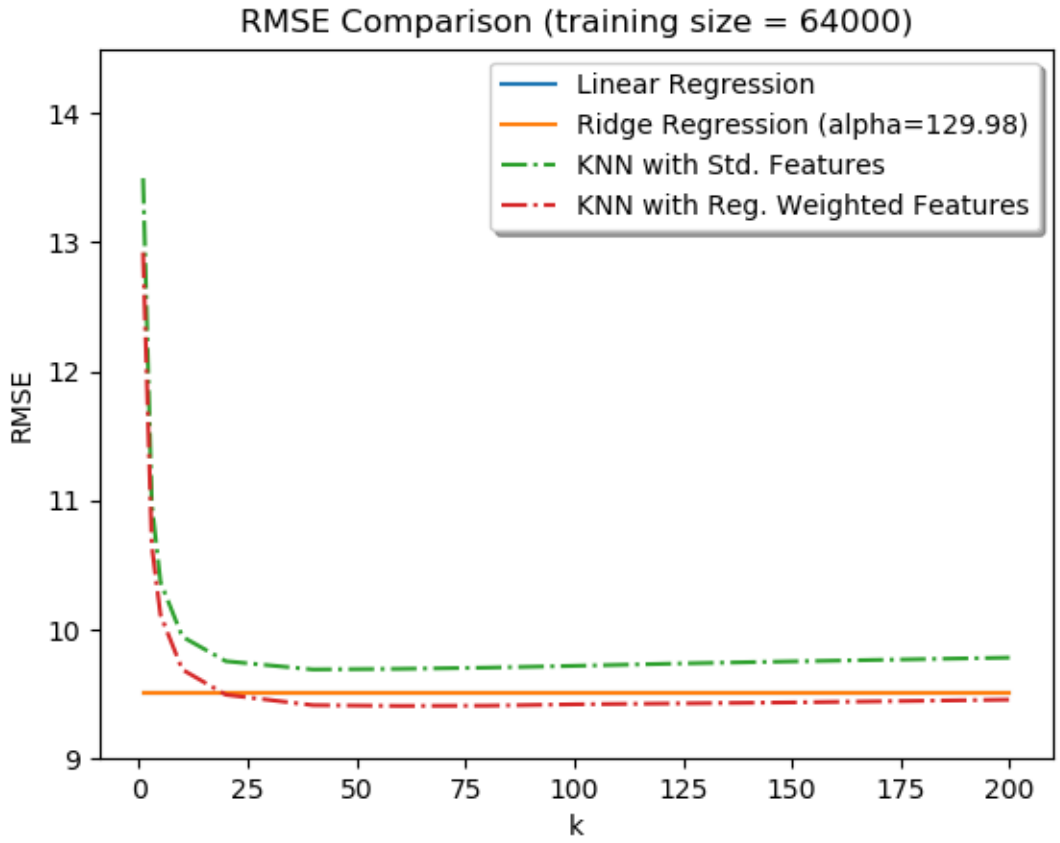
Training = 16,000



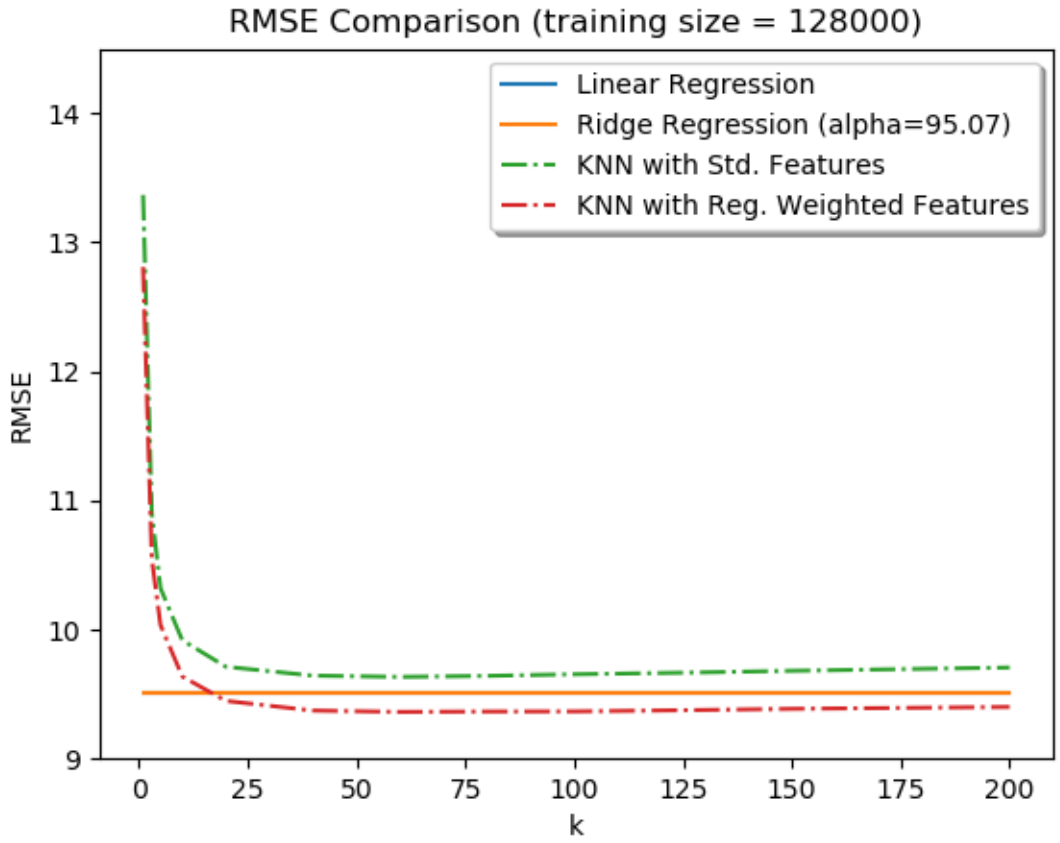
Training = 32,000



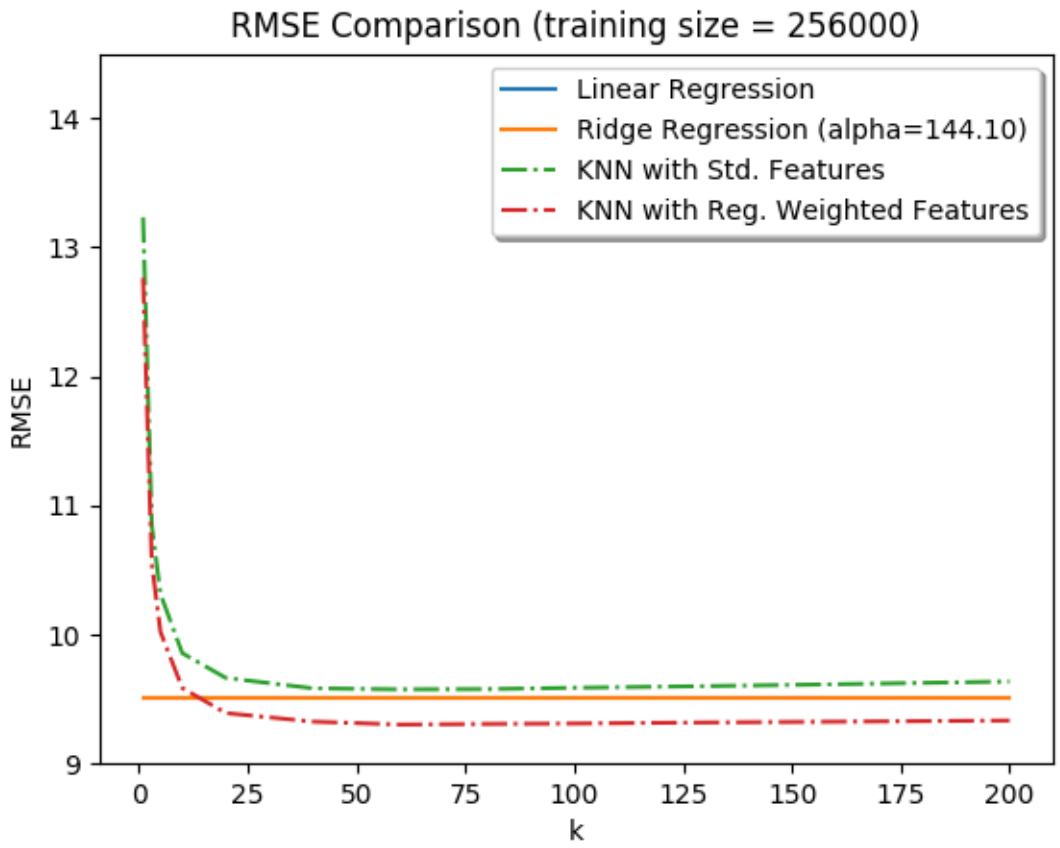
Training = 64,000



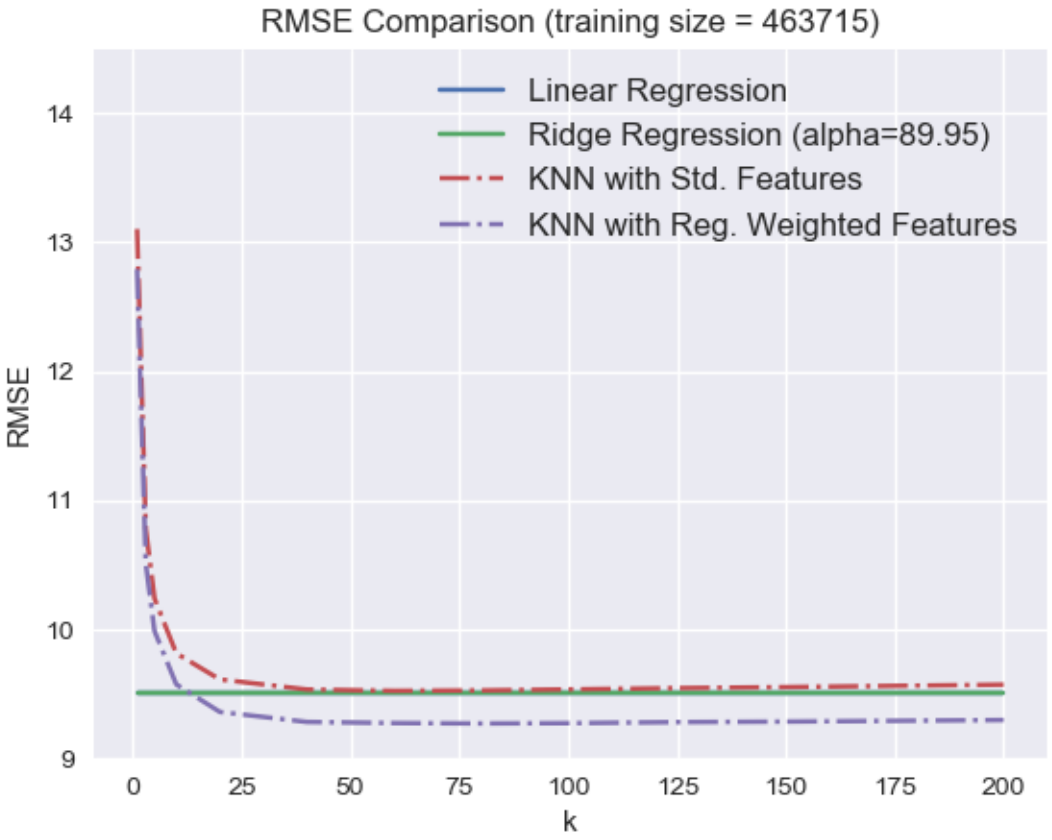
Training = 128,000



Training = 256,000



Training = 463,715



Observations (Part 3)

- Larger datasets do not benefit linear regression models.
- However, kNN regression continue to improve.
- This is an important property for kNN → It can continue benefit from larger dataset.
- However kNN also have a very large drawback
- Any guess?
- It is very slow compared to linear regression
- Often 1,000 times slower compared to linear regression in generating predictions using median-sized training dataset.

Time Comparison

- Intel i5-7600@3.4GHz
- 64GB RAM.
- Single thread.

	Training (mins)	Testing (mins)
Linear Regression	0.05	0.00015
Ridge Regression	0.28	0.00015
kNN (k=40; using Ball Tree)	0.11	68.6

- kNN is 475,333 times slower compared to linear regression in doing predictions.

Performance Summary

- Training data size: 463,715
- Testing data size: 51,630

- RMSE (Linear Regression): 9.510161
- RMSE (Ridge Regression): 9.510161
- RMSE (kNN; Std Features; k=60): 9.524690
- RMSE (kNN; Weighted Features; k=80): 9.271765

- Published RMSE for kNN: 10.20
- See: http://ismir2011.ismir.net/papers/OS6-1.pdf